2/2/2020

**SUBMITTED BY :-**

SHWETA GUPTA

PGPBABI 2019 -2020

**A PROJECT ON -**

**COLD STORAGE CASE STUDY**

**Table of contents**

* Background and objectives – Problem 1 & 2
* Assumptions - Problem 1 & 2

**Analysis problem 1**

* Exploratory data analysis(Q2,Q3)
* Bivariate analysis (Q1)
* Outlier identification
* Assume normal distribution, what is the probability of temperature having fallen below 2 c? (q4)
* Assume normal distribution, what is the probability of temperature having gone above 4 c? (q5)
* What will be the penalty for the AMC company?(q6)

**Analysis problem 2**

* Histogram and boxplot of temperature
* Which hypothesis test shall be performed to check that if corrective action is needed at the cold storage plant? (Q1)
* State the hypothesis, perform hypothesis test and determine p-value (Q2)
* Give your inference (Q3)

**Conclusion**

* **Analysis problem 1**
* **Analysis problem 2**

**Appendix – Source code**

1. BACKGROUND

**PROBLEM 1:**

* Cold Storage started its operations in Jan 2016. They are in the business of storing Pasteurized Fresh Whole or Skimmed Milk, Sweet Cream, Flavoured Milk Drinks.
* To ensure that there is no change of texture, body appearance, separation of fats the optimal temperature to be maintained is between 2 - 4 C.
* In the first year of business they outsourced the plant maintenance work to a professional company with stiff penalty clauses.
* It was agreed that if it was statistically proven that **probability of temperature going outside the 2 - 4 C** during the one-year contract **was above 2.5% and less than 5% then the penalty would be 10% of AMC** (annual maintenance case). **In case it exceeded 5% then the penalty would be 25% of the AMC fee.**
* The average temperature data at date level is given in the file“Cold\_Storage\_Temp\_Data.csv”

**PROBLEM 2:**

* In Mar 2018, Cold Storage started getting complaints from their Clients that they have been getting complaints from end consumers of the dairy products going sour and often smelling.
* On getting these complaints, the supervisor pulls out **data of last 35 days’ temperatures.**
* As a safety measure, the Supervisor decides to be vigilant **to maintain the temperature 3.9 C or below.**
* The data of the last 35 days is in “Cold\_Storage\_Mar2018.csv”
  1. OBJECTIVE

**PROBLEM 1:**

* Find mean cold storage temperature for Summer, Winter and Rainy Season
* Find overall mean for the full year
* Find Standard Deviation for the full year
* Assume Normal distribution, what is the probability of temperature having fallen below 2 C?
* Assume Normal distribution, what is the probability of temperature having gone above 4 C?
* What will be the penalty for the AMC Company?

**PROBLEM 2:**

Assume **3.9 C as upper acceptable value for mean temperature** and at **alpha = 0.1** **do you feel that there is need for some corrective action in the Cold Storage Plant** or is it that the problem is from procurement side from where Cold Storage is getting the Dairy Products.

1. Which Hypothesis test shall be performed to check that if corrective action is needed at the cold storage plant? Justify your answer.
2. State the Hypothesis, perform hypothesis test and determine p-value .
3. Give your inference .
4. ASSUMPTIONS

**PROBLEM 1:**

* Data follows normal distribution for variable Temperature

**PROBLEM 2:**

* Based on Central limit theorem we assume that for the sample drawn for 35 days, the sample means will follow normal distribution if n number of random samples are drawn and as the size of the sample increases and becomes large.

**--------ANALYSIS PROBLEM 1 ----------**

1. EXPLORATORY DATA ANALYSIS
   1. **Environment Set up and Data Import**

Below packages have been used in executing this project

* library(readr) - for reading the csv files
* library(dplyr) – for bivariate analysis and creating cross tables

Further attach function has also been used to easily call variables in the dataset

* 1. **Variable Identification**

**Inferences**

Cold Storage has 365 observations from 4 variables –

* Season factor with 4 levels (initially a character variable corrected to Factor)
* Month factor with 12 levels (initially a character variable corrected to Factor)
* Date - a numerical variable
* Temperature - a numerical variable
* Dataset does not contain any missing values
* The dataset does not seem to have any kind of discrepancies on overserving the first 10 and last rows of the data

**>dim(cold\_storage\_temp)** # for identifying the no of observations and variables

[1] 365 4

**> names(cold\_storage\_temp)** # for identifying the names of variables

[1] "Season" "Month" "Date" "Temperature"

**> str(cold\_storage\_temp)** # for identifying the class of variables

Classes ‘spec\_tbl\_df’, ‘tbl\_df’, ‘tbl’ and 'data.frame': 365 obs. of 4 variables:

$ Season : Factor w/ 3 levels "Rainy","Summer",..: 3 3 3 3 3 3 3 3 3 3 ...

$ Month : Factor w/ 12 levels "Apr","Aug","Dec",..: 5 5 5 5 5 5 5 5 5 5 ...

$ Date : num 1 2 3 4 5 6 7 8 9 10 ...

$ Temperature: num 2.4 2.3 2.4 2.8 2.5 2.4 2.8 2.3 2.4 2.8 ...

**> anyNA(cold\_storage\_temp)** #for identifying any missing values in data

[1] FALSE

**> sapply(cold\_storage\_temp, function(x) sum(is.na(x)))** # for identifying total missing values by each variable

Season Month Date Temperature

0 0 0 0

**> head(cold\_storage\_temp,10)** # first 10 rows

Season Month Date Temperature

1 Winter Jan 1 2.4

2 Winter Jan 2 2.3

3 Winter Jan 3 2.4

4 Winter Jan 4 2.8

5 Winter Jan 5 2.5

6 Winter Jan 6 2.4

7 Winter Jan 7 2.8

8 Winter Jan 8 2.3

9 Winter Jan 9 2.4

10 Winter Jan 10 2.8

**> tail(cold\_storage\_temp,10)** #last 10 rows

Season Month Date Temperature

356 Winter Dec 22 3.3

357 Winter Dec 23 3.0

358 Winter Dec 24 3.7

359 Winter Dec 25 3.2

360 Winter Dec 26 2.7

361 Winter Dec 27 2.7

362 Winter Dec 28 2.3

363 Winter Dec 29 2.6

364 Winter Dec 30 2.3

365 Winter Dec 31 2.9

1. UNIVARIATE ANALYSIS

**Inferences**

* All 3 seasons have equal no of days in data
* The average/mean temperature for full year is **2.963** whereas median temperature **is 2.900**
* The standard deviation of temperature for full year is **0.508**
* Temperature follows a near normal distribution with a slight left skewed data
* Box plot reveals a minor presence of outliers in the variable temperature
  1. **Five Point Summary on the data**

> summary(cold\_storage\_temp) # summary of all variables

Season Month Date Temperature

Rainy :122 Aug : 31 Min. : 1.00 Min. :1.700

Summer:120 Dec : 31 1st Qu.: 8.00 1st Qu.:2.500

Winter:123 Jan : 31 Median :16.00 Median :2.900

Jul : 31 Mean :15.72 Mean :2.963

Mar : 31 3rd Qu.:23.00 3rd Qu.:3.300

May : 31 Max. :31.00 Max. :5.000

(Other):179

* 1. **Overall Mean and Standard Deviation for full Year (Q2 and Q3)**

> mean(Temperature)

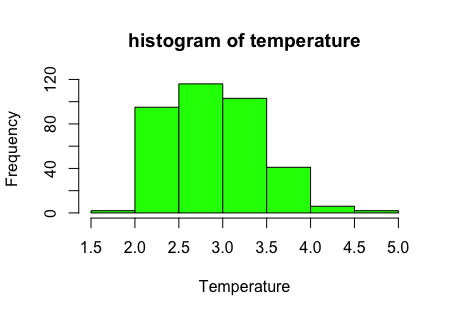
[1] 2.96274

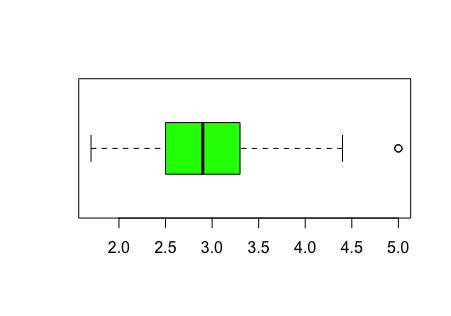
> sd(Temperature)

[1] 0.508589

> var(Temperature)

[1] 0.2586628

* 1. **Histogram and Boxplot of temperature**



**Boxplot of Temperature**

1. BI-VARIATE ANALYSIS

**Inferences**

* Average temperature by season(Q1)
  + - Rainy – 3.04
    - Summer – 3.15
    - Winter – 2.70
* Rainy season has highest fluctuation in temperature compared to Summer and Winter season
  1. **Mean cold storage temperature for Summer, Winter and Rainy Season(Q1)**

> print(temperature\_by\_season )

Season No.of.days average.temp.season median.temp.season sd.season

<fct> <int> <dbl> <dbl> <dbl>

1 Rainy 122 3.04 2.9 0.592

2 Summer 120 3.15 3.2 0.396

3 Winter 123 2.70 2.6 0.401

1. OUTLIER IDENTIFICATION

**Inferences**

* Box plot of temperature reveals an outlier with value more than 4.5
* Two observations found to have temperature more than 4.5
* Both observations are present in rainy season
  1. **Identifying presence of outliers**

> cold\_storage\_temp[cold\_storage\_temp$Temperature>4.5,]

Season Month Date Temperature

<fct> <fct> <dbl> <dbl>

1 Rainy Sep 9 5

2 Rainy Sep 20 5

> print(month\_by\_temperatture)

Month days min.temp max.temp average.temp median.temp

<fct> <int> <dbl> <dbl> <dbl> <dbl>

1 Apr 30 2.5 4.1 3.13 3.15

2 Aug 31 2.3 4.4 3.00 2.9

3 Dec 31 2.1 3.7 2.70 2.6

4 Feb 28 2.6 4 3.23 3.2

5 Jan 31 2.3 3.9 2.70 2.5

6 Jul 31 2.3 4.2 2.96 2.8

7 Jun 30 2.2 4 2.97 2.8

8 Mar 31 2.5 3.9 3.09 3.2

9 May 31 2.6 4 3.17 3.2

10 Nov 30 1.9 3.3 2.60 2.5

11 Oct 31 2.2 3.8 2.80 2.8

12 Sep 30 1.7 5 3.23 3.1

1. ASSUME NORMAL DISTRIBUTION, WHAT IS THE PROBABILITY OF TEMPERATURE HAVING FALLEN BELOW 2 C? (Q4)

**Inferences**

* Probability for temperature to fall below 2 degree C is 2.9%

> probability\_under\_2\_degree = pnorm(2,2.963,0.508)

> probability\_under\_2\_degree

[1] 0.02900189

1. ASSUME NORMAL DISTRIBUTION, WHAT IS THE PROBABILITY OF TEMPERATURE HAVING GONE ABOVE 4 C? (Q5)

**Inferences**

* Probability for temperature to go above 4 degree C is 2.06%

> probability\_over\_5\_degree = pnorm(4,2.963,0.508,lower.tail = FALSE)

> probability\_over\_5\_degree

[1] 0.02060859

1. WHAT WILL BE THE PENALTY FOR THE AMC COMPANY?(Q6)

**Inferences**

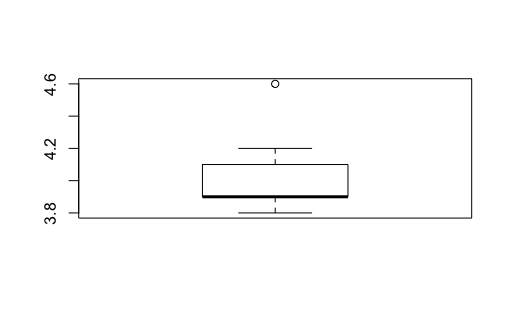
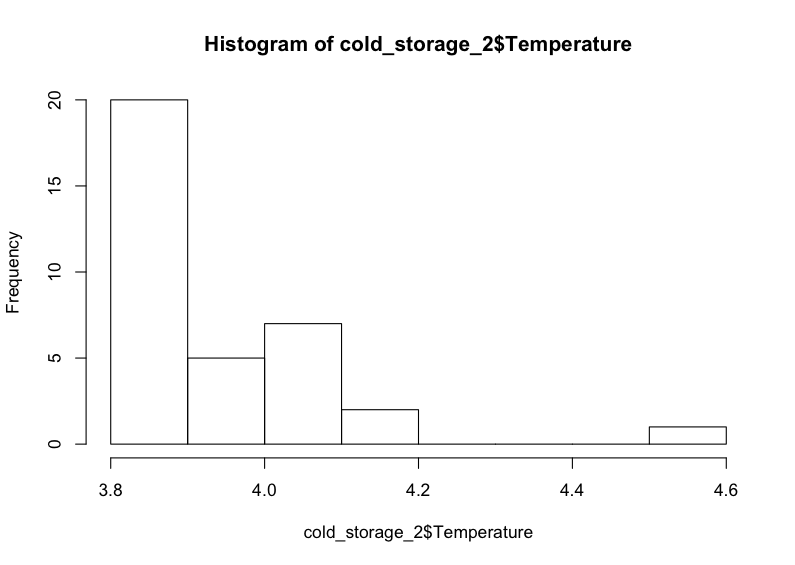
* Total probability of Temperature to go beyond permissible range = Probability of Temperature to go below 2 degree c + Probability of Temperature to go beyond 4 degree c = 2.90% +2.06% = 4.96%
* According to the contract if probability is above 2.5% and less than 5% then the penalty would be 10% of AMC (annual maintenance case). In case it exceeded 5% then the penalty would be 25% of the AMC fee
* **Therefore, since the total probability is within 5% hence the penalty would be 10% of AMC**

**--------ANALYSIS PROBLEM 2 ----------**

**Histogram and Boxplot of Temperature**

**Inferences**

* Temperature data is left skewed with presence of minor outliers in the data

****

Boxplot of Temperature

1. WHICH HYPOTHESIS TEST SHALL BE PERFORMED TO CHECK THAT IF CORRECTIVE ACTION IS NEEDED AT THE COLD STORAGE PLANT?

• The level of significance (Alpha ) = 0.10.

• The sample size , N = 35, which is sufficiently large for a Zstat Test.

• But since the population standard deviation (Sigma) is unknown, we will use **Tstat test.**

• Degree of Freedom: we have N-1 degrees of freedom - 34

•Since the sole purpose of the test is to check whether the temperature is maintained to 3.9 degree or below, we would prefer a **Left tailed T Test.**

1. STATE THE HYPOTHESIS, PERFORM HYPOTHESIS TEST AND DETERMINE P-VALUE

H0: The temperature is >= 3.9 degree

H1: The temperature is <3.9 degree

* Alpha = 0.10
* Confidence interval = 90%
* Degree of freedom = 34
* Xbar = 3.97
* Median of sample = 3.9
* Sd of sample = 0.16
* Mu = 3.90
* Tstat = (xbar – mu) / (sd/ sqaure root of n) =2.75
* Pvalue = 0.004711

**Ttest**

> ttest = t.test(cold\_storage\_2$Temperature,mu = 3.9, conf.level = 0.90,alternative = "greater")

> ttest

One Sample t-test

data: cold\_storage\_2$Temperature

t = 2.7524, df = 34, p-value = 0.004711

alternative hypothesis: true mean is greater than 3.9

90 percent confidence interval:

3.939011 Inf

sample estimates:

mean of x

3.974286

1. GIVE YOUR INFERENCE

* **Since Pvalue (0.0047) is lower than Alpha (0.10) we reject the null hypothesis and conclude that mean temperature is significantly greater than 3.9 degrees**

**CONCLUSION**

**--------ANALYSIS PROBLEM 1 ----------**

* Average temperature by seasons:-
  + - Rainy – 3.04
    - Summer – 3.15
    - Winter – 2.70
* Rainy season has highest fluctuation in temperature compared to Summer and Winter season
* The **average/mean temperature** for **full year** is **2.963** whereas median temperature **is 2.900** indicating minor outlier presence
* The **standard deviation of temperature** for **full year** is **0.508**
* Temperature follows a near normal distribution with a slight left skewed data
* Total probability of Temperature to go beyond permissible range = Probability of Temperature to go below 2 degree c + Probability of Temperature to go beyond 4 degree c = 2.90% +2.06% = 4.96%
* **Therefore, since the total probability is within 5% hence the penalty would be 10% of AMC**

**--------ANALYSIS PROBLEM 2 ----------**

* Since the T Test on sample data proves that mean temperature is significantly higher than 3.9 degree , corrective action needs to be taken at the cold storage facility to bring down the temperatures to desirable levels.
* Although sample mean (3.97) is close to the desirable temperature 3.9 degree and sample median is 3.9 itself, even at lower significance level (Alpha = .10) we failed to reject the null hypothesis indicating that facility indeed needs to take corrective action to better preserve their products.

APPENDIX

(SOURCE CODE)

library(readr) # for reading csv files

cold\_storage\_temp = read\_csv("/Users/shweta/Desktop/PGPBABI/STATISTICAL METHODS AND DECISION MAKING/projects/cold storage/Cold\_Storage\_Temp\_Data.csv")

dim(cold\_storage\_temp) #for identifying the number of observations and variables

View(cold\_storage\_temp)

str(cold\_storage\_temp) # know the class of all variables

head(cold\_storage\_temp,10) # display first 10 rows

tail(cold\_storage\_temp,10) # display last 10 rows

attach(cold\_storage\_temp)

names(cold\_storage\_temp)

summary(cold\_storage\_temp) # summarise all variables

cold\_storage\_temp$Season = as.factor(cold\_storage\_temp$Season) #change the class of variable to factor

cold\_storage\_temp$Month = as.factor(cold\_storage\_temp$Month)

summary(cold\_storage\_temp)

anyNA(cold\_storage\_temp) # identify missing values presence

sapply(cold\_storage\_temp, function(x) sum(is.na(x))) # identify total no of missing values presence by each variable

mean(Temperature) # average

sd(Temperature) # standard deviation

var(Temperature) # variance

hist(Temperature,col = "green",main ="histogram of temperature") # histogram

table(Season,Month)

p = prop.table(table(Season,Month)) # table with proportions

library(dplyr)

temperature\_by\_season = cold\_storage\_temp %>% group\_by(Season) %>% summarise(No.of.days=n(), average.temp.season=mean(Temperature,na.rm = TRUE),median.temp.season = median(Temperature,na.rm = TRUE),sd.season = sd(Temperature,na.rm = TRUE)) # create cross tables

print(temperature\_by\_season )

boxplot(cold\_storage\_temp)

boxplot(Temperature, horizontal = TRUE,col = "green")

cold\_storage\_temp[cold\_storage\_temp$Temperature>4.5,]

table(Month,Temperature)

month\_by\_temperatture = cold\_storage\_temp%>% group\_by(Month) %>% summarise( days =n(), min.temp =min(Temperature),max.temp = max(Temperature),average.temp = mean(Temperature), median.temp = median(Temperature))

print(month\_by\_temperatture)

cold\_storage\_temp$Month = as.factor(cold\_storage\_temp$Month)

class(cold\_storage\_temp$Month)

str(cold\_storage\_temp)

boxplot(Temperature~Month,col = "blue")

by(cold\_storage\_temp,INDICES = Month,FUN = summary) # get summary on all variables by month

IQR(Temperature)

outlier\_range\_positive\_temperature = 3.3+1.5\*0.8 # identify outlier

probability\_under\_2\_degree = pnorm(2,2.963,0.508) # check probability under 2 degree

probability\_over\_4\_degree = pnorm(4,2.963,0.508,lower.tail = FALSE) # check probability over 4 degree

probability\_over\_4\_degree

cold\_storage\_2 = read.csv("/Users/shweta/Desktop/PGPBABI/STATISTICAL METHODS AND DECISION MAKING/projects/cold storage/Cold\_Storage\_Mar2018.csv")

mean\_cold\_storage\_2 = mean(cold\_storage\_2$Temperature)

median\_cold\_storage\_2 = median(cold\_storage\_2$Temperature)

histogram\_cold\_storage\_2 = hist(cold\_storage\_2$Temperature)

boxplot(cold\_storage\_2$Temperature)

mean\_cold\_storage\_2

sd\_cold\_storage\_2 = sd(cold\_storage\_2$Temperature)

sd\_cold\_storage\_2

mu=3.9

n=35

tstat=(mean\_cold\_storage\_2- mu)/(sd\_cold\_storage\_2/(n^0.5)) # calculate tstat

tstat

ttest = t.test(cold\_storage\_2$Temperature,mu = 3.9, conf.level = 0.90,alternative = "greater") # perform 1 tailed t test

ttest

saveRDS(cold\_storage\_temp, "As\_RDS\_cold\_storage\_temp\_2020.RDS")